DOCUMENT RESUME

ED 441 024 TM 030 822

AUTHOR Fisher, William P.; Suttkus, Ramona; DiCarlo, Richard
TITLE Scaling an Introduction to Clinical Medicine Examination.
PUB DATE 2000-04-27
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Higher Education; Medical Education; *Medical Students; *Scaling; *Test Construction
IDENTIFIERS Clinical Competence; *Invariance

ABSTRACT

This paper shows that a substantial degree of invariance can be attained with an examination not explicitly designed to do so, provides an example of how invariance can be demonstrated through plots, and dispels misconceptions concerning the rigidity of the definition of invariance. Responses of 177 examinees to 94 items of a final examination from an introduction to clinical medicine course were obtained. All data analyses involved fitting a dichotomous two-parameter model via unconditional maximum likelihood estimation and were performed using the WINSTEPS program. Even though the examination was designed by persons untrained in educational measurement, the test item orders achieved a substantial degree of invariance across samples of examinees. The measurement reliabilities closely approximated those predicted by Rasch generalizability theory. The plots of the item difficulty estimates from the separate samples and of the examinee ability estimates obtained from the separate subtests support the use of probabilistic conjoint measurement models and the search for invariance. The study shows that a carefully designed test can achieve a high degree of model fit and thus provide a scientific basis for the practical and convenient inferential advantages that follow from the use of sufficient statistics. (Contains 1 table, 8 figures, and 17 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 024

SCALING AN INTRODUCTION TO CLINICAL MEDICINE EXAMINATION

William P. Fisher, Jr., Ph.D.
Ramona Suttkus, Richard DiCarlo, M.D.
LSU Health Sciences Center Biometry & Genetics
1901 Perdido Street
New Orleans, LA 70112

504/568-8083 (Office)
504/568-8500 (Fax)

Paper presented at the 2000 annual meeting of the
American Educational Research Association
in New Orleans, Louisiana
session 37.43
Thursday, 27 April 2000

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☐ This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to
improve reproduction quality.
• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

W. Fisher

TM030822

1

BEST COPY AVAILABLE

OBJECTIVES

Fundamental measurement theory's requirements of parameter invariance, of relatively stable, robust, and unchanging item difficulties across examinees, and examinee abilities across items, strikes many as a rigid and unattainable goal, even when probabilistically formulated. The purposes of this presentation are to 1) show that a substantial degree of invariance can be attained with an examination not explicitly designed to do so; 2) provide a sample exercise of how invariance can be demonstrated via plots; and 3) dispel misconceptions concerning the rigidity of the definition of invariance.

THEORETICAL FRAMEWORK

Hall, Wijsman, and Ghosh (1965) show "that the set of invariant rules based on a sufficient statistic is an essentially complete subclass of the class of invariant rules" (Arnold 1985), where invariance is the stable structure and meaning of a quantitative unit across test/survey questions and respondents. One especially easy to apply set of models for testing for sufficiency and invariance emerged from the works of the Danish mathematician Georg Rasch. Rasch had studied with Fisher in London in the 1930s, and took special care to base his measurement models on the concept of sufficiency, stating a mathematical separability theorem that prescribes the data structures necessary for parameter invariance over samples of persons and items. Rasch focused on sufficiency because, as he (in Wright 1980) later said,

> When a sufficient estimate exists, it extracts every bit of knowledge about a specified feature of the situation made available by the data as formalized by the chosen model. 'Sufficient' stands for 'exhaustive' as regards the feature in question.

> What is left over when a sufficient estimate has been extracted from the data is independent of the trait in question and may therefore be used for a control of the model that does not depend on how the actual estimates happen to reproduce the original data....
>
> The realization of the concept of sufficiency, I think, is a substantial contribution to the theory of knowledge and the high mark of what Fisher did.... His formalization of sufficiency nails down the ... conditions that a model must fulfill in order for it to yield an objective basis for inference.

Fisher (1922) understood sufficiency as a crucial part of the mathematical foundations of theoretical statistics, deploring "the prolonged neglect into which the study of statistics, in its theoretical aspects, has fallen". Michell (1990), echoing Guttman's (1950) comments, interpreted the neglect of the mathematical principles at the foundations of statistical inference as negligence of measurement theory, saying that

> In general psychologists have ... found refuge in quantitative methods that, because they assume more, demand less foundational research as the basis for their application. Methods that always yield a scaling solution, like the method of summated ratings, are almost universally preferred to methods which ... do not produce a scaling solution when they are falsified by the data. Surprisingly, vulnerability to falsification is commonly deemed by psychologists to be a fault rather than a virtue.

This is so even though the Encyclopedia of Statistical Sciences (Arnold 1982-88) states that "most statisticians accept the principle that statistical analysis should depend only on a sufficient statistic." Mining the same vein, S. S. Stevens (1951) wrote that

> The scientist is usually looking for invariance whether he knows it or not.... The quest for invariant relations is essentially the aspiration toward generality, and in psychology, as in physics, the principles that have wide applications are those we prize.

Bachelard (1984), a philosopher of physical science, agrees, saying, "it is in the determination of invariants that the mathematization of the real finds its true justification." It may be that tests of sufficiency and invariance are rare in the psychosocial sciences because most formulations of these tests have been so overly stringent that falsification of the quantitative hypothesis became a virtual certainty (Wilson 1989).

METHODS

All data analyses involved fitting a dichotomous two-parameter (items and persons) model via unconditional maximum likelihood estimation (Wright 1988; Wright & Douglas 1977; Wright & Masters 1982), and were performed using WINSTEPS (Wright & Linacre 1999).

DATA SOURCES

Responses of 177 examinees to a 100-item final exam from a year-long Introduction to Clinical Medicine course offered by an accredited school of medicine were obtained. Six items had been previously removed from the results by the instructor, leaving 94 for analysis. After all the data were analyzed together, another eight analyses were performed. In two of these analyses, the test was divided in half, without determining in which half the items removed by the course director might fall. Each half of the test was then used to produce measures for all 177 examinees, resulting in two sets of measures. In six other analyses, cases were removed from the original analysis in entry order, in three groups of 29 and three groups of 30. Each of the six

groups were then used to produce calibrations for all 94 items, resulting in six sets of calibrations. No item difficulties or person measures were anchored in any analyses.

RESULTS

In the first analysis, for all 94 items and 177 examinees, modeled measurement separation reliability was .82 and calibration separation reliability was .95. Data were 99.8% complete, with the average number of items per examinee being 93.8. The average raw score was about 69, or 73% correct, with a maximum of 90 and a minimum of 42. There were no items or persons with minimum or maximum extreme scores in the overall analysis.

The mean standardized model fit statistics were near 0.0, with standard deviations of less than 1.0, for all nine analyses. Measurement separation reliability varies from .74 to .85 across the six groups of 29 or 30 examinees, with modeled calibration separation reliability ranging from .70 to .77 for these groups. These calibration reliabilities are calculated with the difficulty estimates for items with minimum extreme scores included. There were 2 to 9 items with minimum extreme scores in the six analyses. Excluding these items causes the reliabilities to increase slightly, to a range of .73 to .78.

As shown in Table 1 and Figures 1 to 15, the 15 correlations of the six sets of calibrations range from .81 to .86, with six at the mode of .85, and another three each at .84 and .86. The difficulty estimates for the 2 to 9 items per analysis that obtained the minimum extreme score are included in these correlations; the correlations drop slightly when the extreme items are removed.

One of the six items removed by the course director was in the first half of the test, and the remaining five were in the other half, so the two subtests had 49 and 45 items, respectively. There were no items or persons with minimum or maximum extreme scores in these analyses.

Measurement separation reliabilities were .67 and .71 for the 177 examinees on the two subtests, with calibration separation reliabilities at .95 and .96, respectively. As shown in the figure, the ability estimates of the examinees, as measured by the two different groups of items, correlate .73, as is expected from the measurement separation reliabilities. The regression line and 95% confidence intervals in the figure show that subtest 2 is somewhat easier than subtest one, though the identity line exceeds the width of the confidence intervals only at the extremes, outside of the intended measurement range.

SCIENTIFIC IMPORTANCE

Even though the examination studied in this series of analyses was designed by persons untrained in educational measurement (the course director and instructors), the test item orders achieve a substantial degree of invariance across samples of examinees. Given the number of items and the measurement standard deviation obtained (about .7 logits), the measurement reliabilities closely approximate those predicted by Rasch generalizability theory (Linacre 1993). Even though there was no overlap in examinees across the six samples, the 94 items fell into virtually the same difficulty order for each of them. The correlations of the difficulty estimates probably approach statistical identity after disattenuating for error (Muchinsky 1996; Schumacker 1996), given the rough average of the calibration reliabilities at .76.

The plots of the item difficulty estimates obtained from the separate samples, and of the examinee ability estimates obtained from the separate subtests support the use of probabilistic conjoint measurement models and the search for invariance. The plots may be of special value for audiences unable to follow mathematical logic easily, or who fear that important information

may be lost when a test does not cover all of the items deemed relevant to the main content domains, as with adaptive administration.

Finally, this study shows that a carefully designed test can achieve a high degree of model fit and can thus provide a scientific basis for the practical and convenient inferential advantages that follow from the use of sufficient statistics. Future research will focus on demonstrations of these advantages. For instance, adaptive administration might be simulated from existing data by retrospectively tailoring the test by individualizing it: removing from the analysis all items two logits above or below each examinee's measure. Measures from the simulated adaptive and the entire test would then be compared.

Several successful experiments of this sort could be useful in helping educators think about the curriculum in more abstract and general terms. A more abstract conception of the curriculum and of the abilities measured by tests would be useful because it would free teachers from misplaced concreteness: the sense that every important content area in a subject must be tested, even when a student's probability of success is 100% or 0%. No course of study ever touches on every conceivable variation in the subject matter. Every course is a survey of the subject that samples from a potentially infinite universe of expressions to focus on the ones that seem most relevant or accessible. Students never gain experience with every possible kind of problem in a subject area that might come along, making it all the more important to take care in choosing the questions we ask in assessing their learning.

But in the same way, tests never cover all of the possible course content. For tests to be quantitative measures, it is important, even crucial, to examine the extent to which the items might be said to all belong to the same population of possible items. The studies presented here

show that one way of doing this is to see if different subsets of items produce equivalent measures, and if different subsets of examinees produce equivalent item calibrations.

The positive results support the hope that one day educational measurement practice will achieve the status of those fields deemed scientific, fields that are not merely quantitative, but fields that have so far excelled in the search for invariance that they have coalesced into communities held together by the bond of a common mathematical language. For science is not achieved through the mere use of number or mathematical equations. First and foremost, science is achieved when 1) invariant structures are identified and scaled within individual instrument calibration studies, and 2) those structures are found to retain their invariant character across studies of additional samples of persons and items. What remains then is to agree on and set up systems for maintaining and improving a quantitative unit range and metric for the reference standard to which all instruments measuring the variable are equated.

Table 1

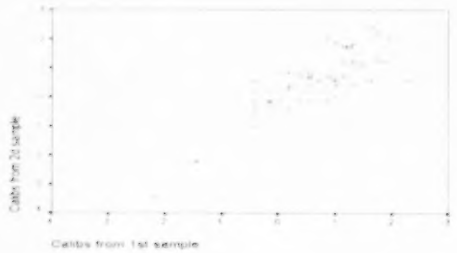
Correlations of Test Items across Six Examinee Subsamples

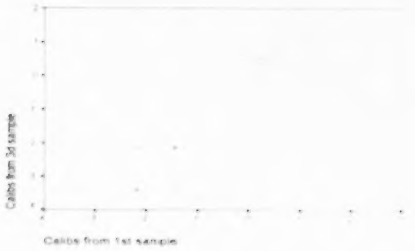
		Calibs from 6th sample	Calibs from 5th sample	Calibs from 4th sample	Calibs from 3d sample	Calibs from 2d sample	Calibs from 1st sample
Calibs from 6th sample	Pearson Correlation	1.000	.854(**)	.843(**)	.828(**)	.848(**)	.857(**)
	Sig. (1-tailed)	.	.000	.000	.000	.000	.000
	N	94	94	94	94	94	94
Calibs from 5th sample	Pearson Correlation	.854(**)	1.000	.854(**)	.837(**)	.858(**)	.851(**)
	Sig. (1-tailed)	.000	.	.000	.000	.000	.000
	N	94	94	94	94	94	94
Calibs from 4th sample	Pearson Correlation	.843(**)	.854(**)	1.000	.811(**)	.860(**)	.842(**)
	Sig. (1-tailed)	.000	.000	.	.000	.000	.000
	N	94	94	94	94	94	94
Calibs from 3d sample	Pearson Correlation	.828(**)	.837(**)	.811(**)	1.000	.845(**)	.852(**)
	Sig. (1-tailed)	.000	.000	.000	.	.000	.000
	N	94	94	94	94	94	94
Calibs from 2d sample	Pearson Correlation	.848(**)	.858(**)	.860(**)	.845(**)	1.000	.853(**)
	Sig. (1-tailed)	.000	.000	.000	.000	.	.000
	N	94	94	94	94	94	94
Calibs from 1st sample	Pearson Correlation	.857(**)	.851(**)	.842(**)	.852(**)	.853(**)	1.000
	Sig. (1-tailed)	.000	.000	.000	.000	.000	.
	N	94	94	94	94	94	94

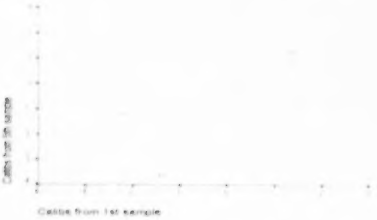
** Correlation is significant at the 0.01 level (1-tailed).

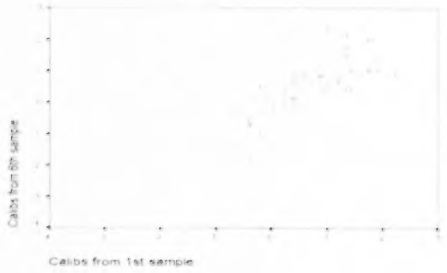
BEST COPY AVAILABLE

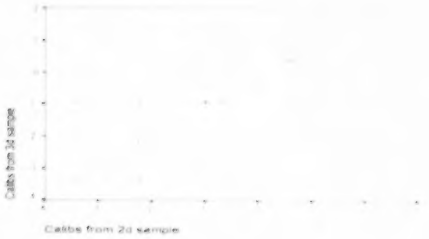
Figures 1 - 8

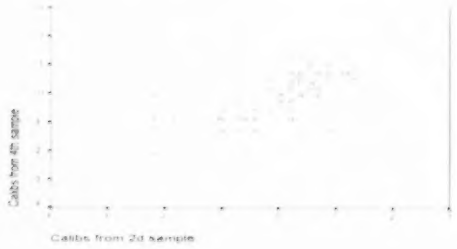






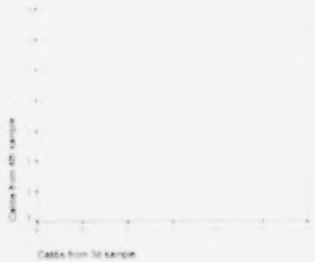


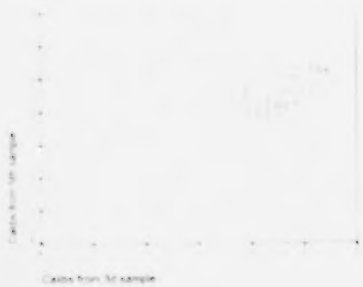


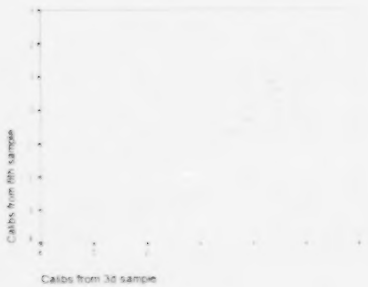


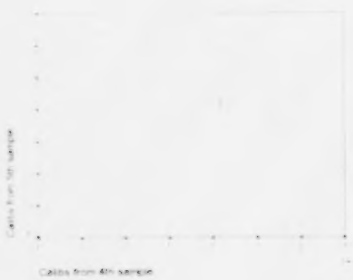
BEST COPY AVAILABLE

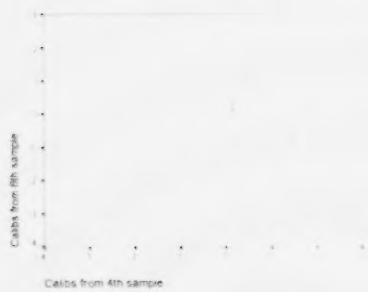
Figures 9 - 15

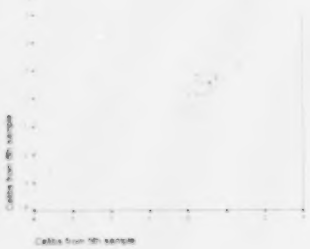












BEST COPY AVAILABLE

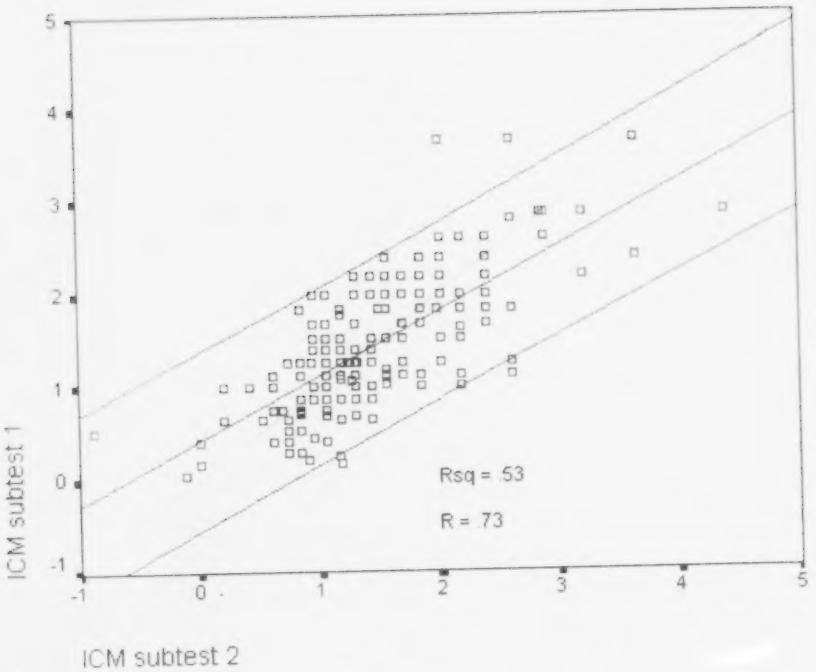


Figure 16.

BEST COPY AVAILABLE

REFERENCES

Arnold, S. F. (1982-1988). Sufficient statistics. In S. Kotz, N. L. Johnson & C. B. Read (Eds.), *Encyclopedia of Statistical Sciences* (pp. 72-80). New York: John Wiley & Sons.

Arnold, S. F. (1985, September). Sufficiency and invariance. *Statistics & Probability Letters, 3*, 275-279.

Bachelard, G. (1984). *The New Scientific Spirit* (Arthur Goldhammer & Foreword by Patrick A. Heelan, Trans.). Boston: Beacon Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A, 222*, 309-368.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer & et al. (Eds.), *Studies in social psychology in World War II. volume 4: Measurement and prediction* (pp. 60-90). New York: Wiley.

Hall, W. J., Wijsman, R. A., & Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *Annals of Mathematical Statistics, 36*, 575-614.

Linacre, J. M. (1993). Rasch generalizability theory. *Rasch Measurement Transactions, 7*(1), 283-284.

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*(1), 63-75.

Schumacker, R. E. (1996, Spring). Disattenuating correlation coefficients. *Rasch Measurement Transactions, 10*(1), 479.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: John Wiley & Sons.

Wilson, M. (1989). A comparison of deterministic and probabilistic approaches to learning structures. *Australian Journal of Education, 33*(2), 127-140.

Wright, B. D. (1980). Foreword, Afterword. In *Probabilistic models for some intelligence and attainment tests, by Georg Rasch* [Reprint; original work published in 1960 by the Danish Institute for Educational Research]. Chicago: University of Chicago Press.

Wright, B. D. (1988, Sep). The efficacy of unconditional maximum likelihood bias correction: Comment on Jansen, Van den Wollenberg, and Wierda. *Applied Psychological Measurement., 12*(3), 315-318.

Wright, B. D., & Douglas, G. A. (1977). Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement, 37*, 47-60.

Wright, B. D., & Linacre, J. M. (1999). *A User's Guide to WINSTEPS Rasch-Model Computer Program, v 2.94*. Chicago: MESA Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

BEST COPY AVAILABLE